CA Journal of Cultural Analytics

Fictionality

Andrew Piper

12.20.16

Peer-Reviewed By: Adam Hammond

Clusters: Genre, NovelTM Special Issue on Genre

Article DOI: 10.22148/16.011

Dataverse DOI: 10.7910/DVN/5WKTZV

> "There is no textual property, syntactical or semantic, that will identify a text as a work of fiction." - John Searle

The distinction between fiction and non-fiction, between a text that is true and one that is not, is one of the oldest on record.[1] Ever since we have been thinking about the act of narration, we have addressed the related meanings of truth and imagination. This is what Aristotle designated as the difference between the communicative use of language (*legein*) and its creative use (*poiein*).[2] For millennia, we have been debating whether there are inherent features of being fictional or whether it is simply a matter of intention, that perhaps there is nothing unique to the language of fictional discourse after all. How do we know when a text is signalling that it is "true" or, by extension, not true?[3] And what might quantity have to tell us about this most elementary of distinctions?

Consider for example the following two passages:

A

> On the short ferry ride from Buckley Bay to Denman Island, Juliet got out of her car and stood at the front of the boat, in the summer breeze. A woman standing there recognized her, and they began to talk. It is not unusual for people to take a second look at Juliet and wonder where they've seen her before, and sometimes, to remember.

B

> Jeff is 24, tall and fit, with shaggy brown hair and an easy smile. After graduating from Brown three years ago, with an honors degree in history and anthropology, he moved back home to the Boston suburbs and started looking for a job. After several months, he found one, as a sales representative for a small Internet provider. He stays in touch with friends from college by text message and email, and still heads downtown on weekends to hang out at Boston's "Brown bars." "It's kinda like I never left college," he says, with a mixture of resignation and pleasure. "Same friends, same aimlessness."

At first glance, these passages share a good deal in common. Both use single proper names (Jeff / Juliet) and markers of place (Boston / Denman Island). They each use a number of pronouns (her, they / he), an occasional adjective (short ferry ride, summer breeze / shaggy brown hair, easy smile), as well as the past and present tense (stood, began, is / is, started, moved). While the second passage uses dialogue, it is not unreasonable to assume that the first text might at some point, too. And both passages seem to offer some kind of psychological underpinning to the description, whether it is Jeff's

[1] I would like to gratefully acknowledge the support of the Social Sciences and Humanities Research Council of Canada (SSHRC) in funding this research.

[2] For a thorough discussion see Käthe Hamburger, The Logic of Literature, 2d ed., trans. Marilyn J. Rose (Bloomington: Indiana UP, 1973), 233 and Gérard Genette, Fiction and Diction, trans. Catherine Porter (Ithaca: Cornell UP, 1993), 1-29.

[3] There is an important distinction to be made between fiction and truth here. This essay is about detecting the qualities that inhere when texts signal to readers their imaginariness or their truth. This is different from detecting whether a text is true or not-true as in the sense of "lie-detection."

malaise as a "sales rep" or the woman who recognizes Juliet in the first passage. Both make claims on our thinking about people and personality.

And yet few readers would have difficulty guessing that passage B is from a work of non-fiction (Michael Kimmel's Guyland) and passage A from a work of fiction (Alice Munro's "Silence"). What is it then that makes this so obvious? We could say, as many in our field do, that fictionality is simply ineffable, that it is a matter of a feeling we get when we read. Who could say what conjures imaginary worlds in readers' minds? Or, on the other hand, we could try to be more precise than my initial description above and quantify as many features as possible about these two passages in order to understand where their salient differences lie at the lexical and syntactic level (Table 1).

Table 1. These results were generated using James Pennebaker's Linguistic Inquiry and Word Count (LIWC) software which explores textual features across 80 different psycho-linguistic dimensions. Only a subset of all features are presented. Values are in percentages.

Feature	Kimmel	Munro	Difference (+/-)
verb	4,81	14,06	9,25
funct	47,12	54,69	7,57
present	1,92	7,81	5,89
WPS	17,33	21,33	4
social	8,65	12,5	3,85
insight	0,96	4,69	3,73
pronoun	5,77	9,38	3,61
past	2,88	6,25	3,37
they	0	3,12	3,12
humans	0	3,12	3,12
ppron	4,81	7,81	3
motion	1,92	4,69	2,77
article	6,73	9,38	2,65
preps	19,23	21,88	2,65
space	8,65	10,94	2,29
tentat	0,96	3,12	2,16
quant	2,88	0	-2,88
bio	2,88	0	-2,88
leisure	2,88	0	-2,88
percept	6,73	3,12	-3,61
incl	11,54	7,81	-3,73
work	3,85	0	-3,85
time	11,54	6,25	-5,29
posemo	5,77	0	-5,77
Quote	5,77	0	-5,77
affect	6,73	0	-6,73
Sixltr	21,15	10,94	-10,21

Looking at these passages in this way, we see not only a broader range of differences between them, but also the strength, and presumably, the significance of these differences. Munro's text now looks more verb-ish than Kimmel's, with more present tense verbs relative to the overall number of words. Her sentences are also somewhat longer, and more pronoun heavy. She uses more articles and prepositions than Kimmel and also vocabularies of insight ("recognize," "remember," "wonder"), tentativeness ("sometimes," "wonder"), human-centeredness ("woman," "people"), and mobility ("car," "ride," "ferry," "boat"). Kimmel, on the other hand, uses many more six-letter words, slightly more commas and periods, more numbers, and a greater vocabulary of affect and work. We might think the literary text would be more affective, but part of Munro's art is submerging feelings so that they are more implicit than explicit ("and wonder where they've seen her before, and sometimes, to remember").

This article is about understanding the differences between fictional and non-fictional texts, the signs that signal to readers

when a story is true or not-true. Rather than look at a single example as I have done above, or even several of them, I will be using a collection of roughly twenty-eight thousand documents, both fictional and non-fictional, to better understand what distinguishes fictional writing from its non-fictional counterpart. Much of my emphasis will focus on the novel as one of the dominant forms of fictional writing from the nineteenth century to the present. Beginning around 1800, when we know the novel began its inexorable quantitative rise to prominence, what makes the novel unique as a form of fictional discourse?[4]

Questions about the nature of fictional speech reached something of a high-point in the 1970s and early 80s, with numerous works in the philosophy of language reflecting on the linguistic cues that marked out a text's truth claims.[5] At stake in this endeavor was an attempt to define and thus potentially control for the reliability of language, the ability to distinguish between the truthful and untruthful content of speech. The work of John Searle became a landmark within this movement, providing a framework that was deeply indebted to a theory of speech acts inherited from J.L. Austin.

For Searle and the community of philosophers gathered around him, the differences between fictional and non-fictional discourse did not depend on the actual content of the speech. Instead, it depended on the combined intentionality of the speaker and receiver, what were known as illocutionary and perlocutionary acts. (We might think of these as frameworks for producing and receiving speech.) As Searle writes, "The utterance acts in fiction are indistingiushable from the utterance acts of serious discourse, and it is for that reason that there is no textual property that will identify a stretch of discourse as a work of fiction."[6] For the philosophers of language, fictionality was not a distinct use of language, but depended on the intentions of both writers and readers and the way those intentions were communicated beyond the boundaries of a text.

For literary theorists of about the same time, literature, as a subset of fictional discourse, similarly came to be defined as an indistinguishable textual entity from the larger category of "writing." Searle's and Austin's speech-act theory was used to generate a more general critique of literary essentialism, that there were unique and potentially timeless qualities to works of literature. As Jaques Derrida wrote, explicitly invoking Searle's philosophy, "No exposition, no discursive form is intrinsically or essentially literary before and outside of the function it is assigned or recognized by a right, that is, a specific intentionality inscribed directly on the social body." Derrida would then continue, "This is the hypothesis I would like to test and submit to your discussion. There is no essence or substance of literature: literature is not. It does not exist."[7] For Derrida and much of the poststructural criticism that followed, literature was the product not of a definable set of features, but a social set of intentions, the frameworks of production and reception that underpinned Searle's speech acts.[8] Translating Searle's position on discursive statements into literary interpretation more generally, literature was seen as liberatory precisely because it was irreducible to any kind of pattern, habit or idiolect.[9]

This article makes a very different claim, one that is based on observing a great deal of instances in which individuals have engaged in fictional or non-fictional writing over the past two centuries. Seen from this perspective, fictionality emerges as a highly legible category at the level of linguistic content ("lexis" in Aristotelian terminology). Such legibility is what allows us to build predictive models that can identify works of fiction with greater than 95% accuracy, and it should be added, that allow human readers to do the same (as in my initial experiment above). Contrary to the beliefs of

[4]While initial, very successful work has been done on the predictability of fiction, here I want to study the particular features that are indicative of fictional writing and what those features have to tell us about the nature of fictionality and the history of the novel in particular. See Ted Underwood, "Understanding Genre in a Collection of a Million Volumes, Interim Report." Figshare. https://dx.doi.org/10.6084/m9.figshare.1281251.v1. Retrieved: 16 27, Feb 20, 2016 (GMT).

[5]Richard M. Gale, "The Fictive Use of Language," Philosophy 66 (1971): 324-340; David Lewis, "Truth in Fiction," American Philosophical Quarterly 15 (1978): 37-46; John R. Searle, "The Logical Status of Fictional Discourse," Expression and Meaning: Studies in the Theory of Speech Acts (Cambridge: Cambridge UP, 1979), 58-76; Hilary Putnam, "Is there a Fact of Matter about Fiction?" Poetics Today 4.1 (1983): 77-82; Benjamin Hrushovski, "Fictionality and Fields of Reference," Poetics Today 5.2 (1984): 227-251; Gregory Currie, The Nature of Fiction (Cambridge: Cambridge UP, 1990).

[6]John R. Searle, "The Logical Status of Fictional Discourse," 68. Everything depends on Searlie's distinction there on "stretch of discourse," itself a stretch of discourse.

[7]Maurice Blanchot, The Instant of My Death and Jaques Derrida, Demeure: Fiction and Testimony (Stanford: Stanford UP, 2000), 28.

[8]Stanley Fish, Is there a text in this class? The Authority of Interpretive Communities (Cambridge: Harvard UP, 1982).

[9]Hardly a thing of the past, this position is now being replayed in the field of "postclassical" narratology, which argues that there are no inherent distinguishing features between fictional and non-fictional narratives. Driven largely by new work in the theory of mind, attention is paid not to the unique features of texts but the cognitive apparatus that is brought to bear on these texts and that is assumed to be common across all kinds of narration. As I will show, not only are fictional narratives highly distinct from non-fictional ones, but their differences are most strongly driven by an attention to sense perception, that is, to a sense of embodiment, making a strict reliance on cognition as narrative's primary framework problematic. For the postclassical narratological position, see J. Alber and M. Fludernik, eds., Postclassical Narratology (Columbus: Ohio State UP, 2010). This new work is driven largely as a response to the "classical" narratological work of Dorrit Cohn, The Distinction of Fiction (Baltimore: JHU Press, 1999) and even earlier Käthe Hamburger, The Logic of Literature (Bloomington: Indiana UP, 1973).

the philosophers of language or different schools of literary critics from poststructuralists to postclassical narratologists, truth claims in language (or their opposite fictionality) are a highly recognizable linguistic aspect of texts. What appeared to be the case at the level of the sentence or "utterance" (what Searle rather vaguely called a "stretch of discourse"), no longer holds when we observe writing at a different level of scale.[10] Placing all of the emphasis on the reader's activity, whether as cognitive predisposition or interpretive freedom, overlooks the powerful and extensive ways that texts mark themselves for their readers according to their fictional nature.

Not only does the research here suggest that fictionality is a highly legible category, but it also appears to have been surprisingly stable for at least two hundred years. While there have undoubtedly been significant changes to the way we tell stories, when we use learning algorithms trained on nineteenth-century texts we can still recognize contemporary novels with an impressive degree of accuracy (about 91%), even if that performance does decrease (history still matters). Indeed, the very features that seem to indicate the uniqueness of novels in the nineteenth-century, for example, appear either to be increasing over time or largely holding steady, even among a diverse range of genres into the twentieth and twenty-first century. While it remains an open question as to the extent to which different genres exhibit these features of fictionality to a similar degree, my initial research suggests that there is a surprising degree of commonality across very diverse types of fictional writing. Such continuity has important and still largely unaddressed implications for how we think about both genre and literary periodization.[11]

Understanding the legibility of fictionality - the extent to which it marks itself off as a cultural practice - has important implications for understanding our own disciplinary practices. Recent emphasis on the historical imbeddedness of creative writing, however valuable, in many ways misses the point of such coherence, the way one of the driving concerns of fiction is to differentiate itself from other kinds of writing. This is not to suggest that fiction is not in some basic sense about the real world, but it does suggest that its center of gravity, what Kleist called a work of art's Schwerpunkt in his essay on the Marionette Theater, is located somewhere else. What I ultimately hope to better understand here then is this center of gravity, the ways in which fiction distinguishes itself as a kind of writing. As we will see, fiction's stability, and the novel's in particular, appears to be based on what we might call "phenomenological investment."[12] The particular nature of the novel's contribution to fictional discourse in the nineteenth century (and beyond) has been its concern not simply with the world around us, but our perceptual encounter with that world, one that includes a great deal of skepticism, prevarication, and negation. To experience the world in the novel is first and foremost to doubt.

To think about fictionality and the novel in these terms inevitably puts pressure on some of the more common scholarly refrains of the recent past. Longstanding questions about the novel's "realism," i.e. the extent to which and the means through which a novel reproduces a given environment, give way in this view to the novel's concerns with a "dramatization of encounter" - both with others and with the world, indeed, with otherness more broadly speaking. Rather than think of the novel as a genre based on its relationship to a knowledge of things, thing-theory serving here as a translation of the rise-of-realism into new terms, according to its quantitatively meaningful components the novel appears far more self-referential in nature, offering us access to the knowledge of knowing. It renders explicit an experience of the hypothetical, a testing relationship to the world.[13] While this may have traditionally been how we have thought about a small subset

[10] Perhaps it is worth elaborating on this point. The philosophers of language were invested in the idea that for any given utterance, there were no defining features that would indicate whether something was true or not in every instance. To use an example from our passages above, the sentence, "A woman standing there recognized her, and they began to talk," does not tell us in any clear way whether it is true or not true, whether it refers to something that actually happened in the world versus something that was imagined. However, when taken in aggregate, language does begin to communicate its intended referentiality. This is one of those cases, as Stanley Cavell liked to point out, where the philosopher's example is always too simple to be able to explain something in the real world.

[11] An open question remains the degree of variability of "fictionality" or the See the work of Ted Underwood that addresses precisely this idea of the stability of genre and Mark Algee-Hewitt et al. which addresses the pre-nineteenth-century heterogeneity of fictional narrative. Ted Underwood, "The Life Cycles of Genres," CA: Journal of Cultural Analytics May 23, 2016: http://culturalanalytics.org/2016/05/the-life-cycles-of-genres/; and Mark Algee-Hewitt, Laura Eidem, Ryan Heuser, Anita Law, and Tanya Llewellyn, "Novel Taxonomies: Prehistories of Genre in the Eighteenth Century, CA (forthcoming). In a non-computational vein, Thomas Pavel argues that the unity of the nineteenth- and twentieth-century novel is an intentional aesthetic contrast to the early modern prioritization of"sub-genres." "Early modern narrative culture emphasized the differences between subgenres, while later forms of the novel are the result of multiple attempts to blend these subgenres together." Thomas Pavel, The Lives of the Novel: A History (Princeton: Princeton UP, 2015), 10.

[12] This argument picks-up on Dorrit Cohn's earlier, pre-computational belief that "fiction is recognizable as fiction only if and when it actualizes its focalizing potential." Dorrit Cohn, The Distinction of Fiction, 25.

[13] In this, I see my findings aligning closely with the "possible worlds theory" that was influential in narrative theory in the 1990s. See Thomas Pavel, Fictional Worlds (Cambridge: Harvard UP, 1989); Marie Laure-Ryan, Possible Worlds, Artificial Intelligence, and Narrative Theory (Bloomington: Indiana UP, 1992); and Ruth Ronen, Possible Worlds in Literary Theory (Cambridge: Cambridge UP, 1994). It also lends support for the more recent new historical work of John Bender that emphasizes the relationship between the coeval rise of the novel in the eighteenth century and scientific experimentation. See John Bender, Ends of Enlightenment (Stanford: Stanford UP, 2012).

of modernist experiments, it is significant that this insight holds even across the most canonical "realist" novels of the nineteenth century.

This article thus suggests that certain canonical positions within the history of novel scholarship need to be rethought or at least subject to revision in light of an emerging computational understanding of the novel. Whether it is Catherine Gallagher's argument about the novel's ambiguous relationship to its own fictionality; the poststructural investment in literature's negativity, as when critics speak of the novel as the "genre of no genre"; thing-theory's emphasis on what Elaine Freedgood has called the "denotative, literal, and technical" language of the novel; or Ian Watt's still-influential position on the novel's referentiality, as when he writes, "It would appear then that the function of language is much more largely referential in the novel than in other literary forms"; computation presents a very different portrait of the novel's importance as a type of fictional discourse.[14] The point is not that these positions are unfounded - it is certain that for some novels these ways of being may indeed be predominant just as it is certain that for many novels these ways of being may be operative some of the time.

But if we try to understand what makes the novel stand out from other types of ostensibly true writing or even other types of fictional texts - if we try to generalize about the novel as a genre - then at least since the turn of the nineteenth century we are seeing something altogether different at stake. According to the research I will present here, the novel's mattering is not primarily grounded in its positive representation of the world, that is, in its mimetic utility, its ability to simulate something (as in seventeenth-century debates about vraisemblance). Nor is it grounded in a kind of post-structural negativity - the novel is unrecognizable as a distinct and stable category, a reflection of literature's more general negative capability. Rather, the novel can best be described through its investment in the negation of the certainty of its own worldliness. It is grounded in an appeal to encounter rather than reality. In doing so, it is precisely the referentiality of language that is being bracketed in the novel, not ambiguously, but programmatically, even in novels that are widely considered to be the most realistic.

Prediction and Description

This article will be using a combination of what are called predictive and descriptive methods. Predictive models, such as those employed in the process of machine learning, are important because they allow us to engage in the process of classification, of what it means to define a group of texts as a coherent entity and to understand the degree of coherence according to certain predefined conditions.[15] Predictive models allow us to say with how much certainty we can identify texts that belong to a specific group and under what criteria. The more certainty there is, the more cohesive the category is thought to be.

Descriptive models, on the other hand, are useful because they allow us to qualify distinctive features of one group when compared with another without engaging in the act of classification. They can tell us which features are distinctive of one group versus another, but they do not do so in order to make claims about the overall uniqueness of that group. Instead of defining a text or group of texts - the novel is X or the novel is this predictable - these qualities help describe the behavior of a group according to more individualized criteria. This too is valuable because it allows us to understand the components that make one group different from another but that do not necessarily lead to categorical difference. Explaining predictions - how a computer arrived at an estimate about which class a text belongs to - is quite challenging. Explaining individual differences is far more straightforward. It is their combination, I would argue, that allows us to think both categorically - about the relative coherence of writing under certain conditions - as well as qualitatively about

[14] As Catherine Gallagher writes, "If a genre can be thought of as having an attitude, the novel has seemed ambivalent toward its fictionality - at once inventing it as an ontological ground and placing severe constraints upon it." See Catherine Gallagher, "The Rise of Fictionality," The Novel, vol. 1, ed. Franco Moretti (Princeton: Princeton UP, 2006), 336-363; Elaine Freedgood, "Denotatively, Technically, Literally," Representations 125 (Winter, 2014): 1-14; Frances Ferguson, "Now It's Personal: D.A. Miller and Too-Close-Reading," Critical Inquiry 41 (Spring 2015), 527; and Ian Watt, The Rise of the Novel: Studies in Defoe, Richardson, and Fielding (Berkeley: California UP, 2001). It should be added that Freedgood's emphasis is by no means normative in her identification of the importance of the novel's technical vocabulary; she simply wants us to attend to this overseen dimension because of the way it expands the archive of reading. As I will show at the close of this chapter, this attention to factuality within the novel is something that computational approaches are well suited to address. It is also worth noting how these positions all focus on an earlier time-frame than I am exploring here, and yet their time-frames still serve as the basis of normative arguments about the novel more generally ("our kind of fiction" in Michael McKeon's words). An open question is whether these ambiguities of fictionality or sense of heightened referentiality look that way because of what is happening in the seventeenth-century novel, that is, whether these arguments are based on a different kind of shift occurring around the turn of the eighteenth century that is no longer operative by the nineteenth.

[15] For a fuller discussion and use of predictive models, see Ted Underwood, "The Life Cycles of Genre."

the specific aspects of writing regardless of drawing definitive boundaries around things. Description is in many ways much closer to the traditional task of literary criticism than prediction.

The data that I will be using for this article has been selected to understand the nature of fictionality across different types of writing.[16] The aim is to see if the results here hold across time, different languages, and different sample sizes. Overall, the data consist of ~28,000 documents, dating from the late eighteenth century to the early twenty first written in both English and German. The collections contain different kinds of fictional and non-fictional writing, including novels, histories, philosophy, advice books, novellas, fairy tales, and classical epics translated into prose among other kinds of writing (though not including encyclopedias or cook-books). Together, the texts can be grouped into four principal categories.

The first collection represents a canonical set of nineteenth-century writing of ~600 documents in both German and English curated by my lab. These include the best-known novels from the period as well as well-known non-fiction, including philosophy, essays and histories. These texts have been sufficiently cleaned so that they are subject to minimal transcription errors and also broken out by point of view so that we can control, for example, for third person novels when comparing to historical narratives.

The second collection is a much larger sample of nineteenth century writing in English consisting of 21,158 documents, both fictional and non-fictional, that are drawn from Ted Underwood's research using the Hathi Trust archives.[17] This group, whose contents are much less well understood, allows us to test our results between a canonical subset and a much broader group of writing from the same period. The third component consists of a collection of ~6,500 novels drawn from both the Stanford Literary Lab's nineteenth-century novel collection and the Chicago Text Lab's collection of twentieth-century novels. Together, these collections allow us to examine diachronic shifts in the novel's vocabulary. Finally, the fourth component consists of a collection of 800 contemporary novels and non-fiction published within the past decade curated by my lab.[18] This gives us some traction on the extent to which the effects we are seeing in the past continue to hold in the present. Table 2 provides an overview of the different components that will be used and the respective number of documents.

Table 2. Data Overview

Collection	Key	Description	# Documents
19C Canon	EN_FIC	English Fiction	100
	EN_NOV	English Novels	100
	EN_NOV_3P	Eng. Novels 3-Person	107
	EN_NON	English Non-Fiction	100
	EN_HIST	English Histories	85
	DE_NOV	German Novels	100
	DE_NOV_3P	Ger. Novels 3-Person	110
	DE_NON	German Non-Fiction	100
	DE_HIST	German Histories	75

`HATHI\_FIC Hathi Trust Fiction 9,426`

Hathi Trust (19C) HATHI_NON Hathi Trust Non-Fiction 11,732 HATHI_TALES Hathi Trust Fiction Minus Novels 428

1790-1990 STAN_KLAB English Novels 6,421

[16] Data and code for this project is located here: Piper, Andrew, 2016, "Fictionality", doi:10.7910/DVN/5WKTZV, Harvard Dataverse, V1

[17] The designations of fiction and non-fiction are derived from Ted Underwood's collection, which is located here: https://dx.doi.org/10.6084/m9.figshare.1281251.v1. All duplicate titles were removed, all documents with the word stems for "essays," "tales," "scenes" or "stories" in either the genre or title fields were removed, and only those works with an 89% or better chance of containing more than 80% pages of fiction were chosen.

[18] Similar to the .txtLAB nineteenth-century collection, the documents here represent a canonical representation of fiction and non-fiction of the past ten years, meaning they have passed through some kind of filter, whether it is the New York Times Book Review, a literary prize competition short list, or appeared on various bestseller lists of platforms like Amazon.com or the New York Times. For a more thorough review of the data set and its insights into contemporary forms of social value surrounding the novel, see Andrew Piper and Eva Portelance, "How Cultural Capital Works: Prizewinning Novels, Bestsellers, and the Time of Reading," Post-45 05.10.2016: http://post45.research.yale.edu/2016/05/how-cultural-capital-works-prizewinning-novels-bestsellers-and-the-time-of-reading/.

CONT_NOV	Contemporary Novels	200

Contemporary CONT_NOV_3P Cont. Novels 3-Person 210 CONT_NON Contemporary Non-Fiction 200 CONT_HIST Contemporary Histories 200

The features that I will be exploring will be drawn largely from the LIWC software designed by James Pennebaker. For those not familiar with LIWC, it is a tool developed in the social sciences for studying large text collections. It consists of eighty different features that range from the identification of syntactic and grammatical features like the use of punctuation, prepositions, verb tense, and pronouns to higher-level cognitive phenomena like social, perceptual, or emotional processes.[19] These are dictionaries that have been tested and validated on human subjects, the results of which are available for review.[20] As with all lexicon-based approaches, there are open questions as to the semantic coherence of a given category. Are all of the words in the "insight" dictionary really indicative of moments of cognitive insight in novels? Or do all instances of "I" mean the same thing?[21] The interpretation of these results thus needs to be handled with a good deal of caution. In particular, care needs to be taken in how the categories are understood as categories, with emphasis given to assessing the semantic coherence of these categories within the novels. When we drill down into the individual features, what do we find? As we will see, much of my emphasis will be on the less semantically ambiguous categories, from punctuation and pronouns to verbs of sensory experience or cognitive prevarication.

At the same time, it is also important to emphasize the benefits of such lexicon-building approaches for the computational study of literature. Unlike the problems faced by topic modeling or word embeddings, where collections of words are discovered and labels applied after the fact, here we start with prior assumptions about linguistic categories and test their presence within given text collections. The externality of the words from the collections they are meant to study allows one to test beliefs independently of the collections themselves. While neither approach is perfect, in both cases we are moving between individual words and the ideas those words are thought to embody. What is ultimately at stake is the confidence of a model to approximate some underlying textual phenomenon.[22] The key is making as explicit as possible how we move between these different levels of analytical scale, that is, how we connect the conceptual, the lexical, and the theoretical.

LIWC can thus give us an initial range of intuitively meaningful interpretive categories to build on as well as the lexica upon which those categories may in part be based. One of its principle advantages is that the feature sets can be shared even on proprietary data, as I have done here. Nevertheless, the categories should not be taken at face value, but looked into, as with all semantic fields. Because the dictionaries are transparent in LIWC, users can refine or alter the dictionaries as they see fit, as I have done on occasion here.[23] They can also be combined with other, more customized features, as I have also done here, for example by looking at word classes using a tool like WordNet. While future work will want to continue to expand and refine these kinds of feature sets, the LIWC collection can serve as a useful starting point for any supervised approach to understanding the quantitative dimensions of texts.

The Coherence of Fictionality

The question that I want to begin with is, How coherent is fiction as a type of writing? Are there indeed no syntactic or semantic properties, as Searle contends, that allow us to predict whether something is intended fictionally? Is fictionality exclusively a function of communicative context, the intentionality of the writer and the belief-system of the reader? Or are there features that appear with a high degree of regularity in fictional texts that do not appear in non-fiction such that even a computer can make accurate guesses as to the nature of the text?

[19] For those interested in studying the dictionaries used by LIWC, see their language manual: http://www.liwc.net/LIWC2007LanguageManual.pdf.

[20] See Yla R. Tausczik and James W. Pennebaker, "LIWC and Computerized Text Analysis Methods," Journal of Language and Social Psychology 29.1 (2010): 24-54.

[21] For a thoughtful and practical review of lexicon-driven research, see H. A. Schwartz, J. C. Eichstaedt, L. Dziurzynski, M. L. Kern, E. Blanco, S. Ramones, M. E. P Seligman, and L. H. Ungar, "Choosing the Right Words: Characterizing and Reducing Error of the Word Count Approach." In Proceedings of SEM-2013: Second Joint Conference on Lexical and Computational Semantics, Atlanta, Georgia, USA (2013): 296-305. For a comparison of lexicon versus machine-learning approaches to text analysis, see Andrew Piper and Eva Portelance, "How Cultural Capital Works."

[22] On modeling and computational hermeneutics, see Andrew Piper, "Novel Devotions: Conversional Reading, Computational Modeling and the Modern Novel," New Literary History 46.1 (2015): 63-98.

[23] All dictionaries and modifications are included in the dataset released with this article:

In order to answer these questions, I will use the process known as machine learning to see how accurately a computer can predict a text's given class (I will be using the learning algorithm known as a support vector machine (SVM), which is applied in many text classification scenarios).[24] For those not familiar with this process, a learning algorithm is "trained" on features found in a set of documents for which the classes are already known and then asked to predict which class a group of texts belong to that it has not seen. In this case, I train the algorithm using the LIWC features discovered in a given set of documents and use a process of 10-fold cross-validation to make predictions on whether a document is a work of fiction or non-fiction. What this means in practice is that I randomly divide the corpus into a 90-10 split ten times, where 90% of the documents are used to train the algorithm and the unseen 10% are used to test its reliability. (The folds function in the kernlab package ensures that the folds are equally balanced between the two categories.) Doing this 10 times allows us to gain a full view of all of the documents in the collection as each document has an opportunity to be in the test set. Table 3 presents the results of this experiment, showing which two data sets were compared and the average accuracy score of the predictions on the unseen data.

As we can see in Table 3, not only are the differences between fiction and non-fiction robust across time and languages, but we can use models built in one time period to strongly predict those of another. While there is a clear drop in performance when we use nineteenth-century models to predict twenty-first-century novels, we can still see a relatively high degree of performance at work here (around 91% accuracy). There appears to be a notable degree of diachronic stability to fictional discourse over the past two centuries. Indeed, as we will see below, when we examine features that are more indicative of novelistic writing in particular as a subset of fictional discourse, we generally see these features increase over time. The trans-temporal stability of the novel is complemented by an increase in certain types of novel-specific vocabulary that can be traced back to the nineteenth century.

Table 3. The Predictability of Fiction

Corpus1	Corpus2	Accuracy (F1)	No. Docs
Fiction (EN_FIC)	Non-Fiction (EN_NON)	0,94	100/100
English Novel (EN_NOV)	Non-Fiction (EN_NON)	0,96	100/100
German Novel (DE_NOV)	Non-Fiction (DE_NON)	0,95	100/100
English Novel 3P (EN_NOV_3P)	History (EN_HIST)	0,99	95/86
Germ Novel 3P (DE_NOV_3P)	History (DE_HIST)	0,99	88/75
Cont. Novel (CONT_NOV)	Non-Fiction (CONT_NON)	0,96	193/200
Cont. Novel 3P (CONT_NOV_3P)	History (CONT_HIST)	0,99	210/200
19C Fiction (HATHI) (Trained)	20C Novel (CONT) (Tested)	0,91	20,344/393

This table presents the results of classification tests using an SVM classifier with the LIWC feature set. In the final example, I trained the classifier on the fiction and non-fiction of the Hathi Trust data set and then ran it against the collection of contemporary novels reviewed in the New York Times and popular contemporary non-fiction.

The Phenomenology of the Novel

If fiction is so predictable, what are the features or combinations of features that make it so? While there are many ways one might want to approach this question, here I rank features by the increase of their median frequency in fiction versus non-fiction and use a non-parametric Wilcoxon Rank Sum Test to indicate statistical significance. The value of comparing medians is that it preserves information about the overall distribution of a feature in a given sample. Rather than lump all works of fiction together into a single bin, where some works may have significantly higher amounts of a feature than others and thus skew the results in their favor, the median identifies the mid-point of a given category. Second, where some significance tests are designed to avoid the importance of lower-frequency features, this test does not make that assumption. It looks exclusively at the ratio between the two populations to understand how much more frequent a given feature is relative to its overall occurrence. Low-occurring features that increase a great deal are thus considered more

[24] The kernlab package was used in R. All runs use a Gaussian kernel ("rbfdot"). A useful introductory text to machine learning is Brett Lantz, Machine Learning with R (Birmingham, UK: Packt, 2013).

important than high-occurring features that increase by a smaller amount.[25]

To begin, let me review the overall structure of the tables used here to better understand what they can and cannot tell us (Table 4). The left most column ("Feature") lists the features as defined by LIWC. Some are extremely straightforward ("exclamation" refers to the percentage of exclamation marks), while others are more nuanced. "Family," for example, refers to a dictionary of words all related to family members, while "social" relates to words having to do with social experience, which for example can include pronouns (a choice that effectively duplicates the pronouns categories because they are so much more common than other words). The former is arguably much more straightforward than the latter and thus we need to be cautious when we encounter a dictionary that is more semantically ambiguous (though even a single word like "you" may have different kinds of functions in novels). The second column ("Category") lists the category to which the feature belongs, a slightly more general framework for understanding the individual features. The next two columns (Fiction %, Non-Fiction %) present the median frequency of that feature in each corpus as a percentage of all words. This allows us to see which features are more prevalent relative to other features.

Because percentages are somewhat opaque in terms of a reader's experience, I will generally be translating these numbers into page and work equivalents in the discussion that follows. This allows us to imagine our way into a reader's experience and surmise which features occupy more of a reader's attention. Exclamation marks, for example, comprise on average about .45% of a given work of fiction in the nineteenth century. If we assume an average novel length of about 100,000 words (or 500 words per page across 200 pages), this means that there is one exclamation mark for about every 200 words, or 2-3 per page, or roughly 500 total per novel. Personal pronouns, on the other hand, occur about 10% of the time in fiction, which means once every 10 words, or 50 times per page (and 10,000 times per novel).

The fifth column, "Ratio," lets us see how much more prevalent the feature is in one collection over another. Exclamation marks appear almost ten times as often in fiction than in non-fiction. This is a massive difference, but we are still only talking about something that occurs relatively infrequently compared to other features. Personal pronouns, on the other hand, only appear a little more than two-times more often in fiction (still a very large difference), but this increase is based on a much larger linguistic aspect of texts. Two times as many pronouns means roughly 5,000 more pronouns per work or about 25 more per page. While I privilege ratio here in my interpretation of the results, we will want to keep our eye on both of these aspects, from the overall prevalence of the feature to the relative increase from one population to another.

Beginning with the baseline comparison of fiction and non-fiction writing using both our canonical sample and the larger collection of Hathi Trust writings from the nineteenth century (Table 4), we see how the features that are most indicative of fictionality are driven by dialogue - exclamation marks, question marks, quotation marks, first and second person pronouns like "I" and "you," assent words like "yes," "okay," and "oh," and finally the word "said" (which is labeled as an auditory verb by LIWC).

Importantly, we also see very strong alignment between the nineteenth-century sample and the larger population of Hathi Trust documents, with some notable exceptions around the "social" category and potentially "family," "home," and "ingestion." If we compare these groups directly, we see that only "family" and "ingestion" are somewhat inflated in the canonical sample (by about 10-15%).[26] In other words, while there are interesting variations that are worth exploring, on the whole, the smaller sample does a good job of capturing the same information as the larger collection. Taken together, these features suggest a relatively unambiguous way in which fictional writing has a uniquely dialogical construction when compared with non-fiction. While this may not be "news," it does help us build a taxononomy of the distinctions that make this kind of writing socially significant. Imagining people talking to each other appears to be one of fiction's primary cultural functions.

Table 4. Fiction v. Non-Fiction
19C Canon (English)

Feature	Category	Fiction	Non-Fiction (%)	Ratio	Sample rank	Hathi rank	p-value
exclam	linguistic	0,4	0,04	10	1	2	***
you	linguistic	1,34	0,16	8,34	2	1	***

[25]It is entirely possible to reverse this assumption and privilege features that are more prevalent overall, something that becomes valuable when assessing individual words. This is the approach I use in Table 11 for example. Individual words can have such low frequencies that favoring words with higher counts assures that one is finding more "important" or perhaps "relevant" vocabulary, i.e. less random vocabulary. The crucial point is that the outcomes are determined by the initial assumptions used in the model.

[26]See LIWC_Comparison_EN_FIC_v_HATHI_FIC.csv.

Feature	Category	Fiction	Non-Fiction (%)	Ratio	Sample rank	Hathi rank	p-value
q-mark	linguistic	0,41	0,08	5,13	3	6	***
I	linguistic	2,41	0,49	4,92	4	7	***
quote	linguistic	2,59	0,65	3,98	5	5	***
assent	social	0,12	0,03	3,83	6	4	***
family	social	0,56	0,16	3,58	7	10	***
hear	perception	1,15	0,38	3,01	8	9	***
shehe	linguistic	4,86	1,9	2,56	9	8	***
ppron	linguistic	10,73	5,01	2,14	10	14	***
body	biological	1,12	0,55	2,06	11	15	***
home	social	0,57	0,28	2,02	12	19	***
friend	social	0,19	0,1	2	13	11	***
sexual	social	0,18	0,09	2	14	12	***
see	perception	1,15	0,59	1,94	15	13	***
percept	perception	3,03	1,61	1,88	16	16	***
past	linguistic	6,11	3,37	1,82	17	20	***
feel	perception	0,66	0,39	1,68	18	18	***
social	social	13,25	8,05	1,65	19	33	***
ingest	biological	0,29	0,18	1,61	20	31	***

*The top twenty features with the greatest increase in fiction compared to non-fiction. Values represent median percentage for a given feature. The following code is used to represent p-values: < 0.0001 = ***, < 0.001 = **, < 0.01 = *.*

Indeed, imagining people as people may be fiction's most important role. If we remove dialogue from the sets above, including the pronominal expressions that accompany them (she said, he cried, etc),[27] we see how third-person pronouns emerge as one of the strongest indicators of fictionality along with references to family members and bodies (Table 5). There is over a three-fold increase in the average number of she/he pronouns in fiction versus non-fiction outside of dialogue, with just these two words alone accounting for more than five-percent of all words in the text (or roughly 5,000 instances for a medium-length novel).

This is especially remarkable considering that on average works of history, for example, use considerably more proper names than works of fiction (an estimated more than 2x as many).[28] The lower number of people in fiction is compensated for by a more expanded durational existence on the page for which pronouns become key linguistic markers. People seem to have more extended identities in fiction, though this is not necessarily to be confused with a more "expansive" identity, i.e., one that is more semantically rich. The pronominal frequency of characters is not the same as the linguistic diversity surrounding these characters. Nevertheless, this gives us a first indication of the ways in which fiction performs the process of identification as a repetitive and extensive act of naming the same person.

The prevalence of family and friend vocabulary in fiction also suggests what type of people are more distinctive of the genre, just as the setting of home gives us an idea of where they are most active. Broadly speaking, when we read fiction in the nineteenth century, what is novel, i.e. different from other kinds of texts that purport to be about real things, is a focus on family and the familiar. Travel, adventure, work - these can be experienced elsewhere in ways that documentation of family life and the extended agency of individuals cannot.

Table 5. Fiction v. Non-Fiction with Dialogue Removed
19C Canon (English)

Feature	Category	Fiction (%)	Non-Fiction (%)	Ratio	p-value
family	social	0,53	0,15	3,53	***
exclam	linguistic	0,07	0,02	3,5	***

[27] In addition to removing dialogue, 200 verbs of communication were removed from the texts and any personal pronouns that appeared within +/- 1 word (I said, said she, etc.).

[28] Using the NLP package in R, the mean number of named persons in a novel was 10.58 where it was 23.94 for history. See NER_19C_EN.csv and NER_19C_History.csv. See Script 3.1.

Feature	Category	Fiction (%)	Non-Fiction (%)	Ratio	p-value
shehe	linguistic	5,92	1,92	3,09	***
body	biological	1,28	0,5	2,56	***
home	social	0,71	0,29	2,43	***
sexual	social	0,18	0,08	2,19	***
see	perceptual	1,26	0,59	2,14	***
hear	perceptual	0,65	0,31	2,13	***
past	linguistic	6,49	3,11	2,09	***
feel	perceptual	0,77	0,37	2,07	***
friend	social	0,18	0,1	1,89	***
percept	perceptual	2,81	1,49	1,89	***
anx	affective	0,47	0,26	1,81	***
ppron	linguistic	8,73	4,9	1,78	***
bio	biological	2,23	1,27	1,76	***
ingest	biological	0,3	0,18	1,67	***
social	social	11,81	7,69	1,54	***
sad	affective	0,6	0,39	1,53	***
motion	relative	2,23	1,49	1,5	***
assent	oral	0,03	0,02	1,5	***

The stakes of this attention will become even clearer when we focus on a particular type of fiction (novels with external narrators) and a particular type of non-fiction (history writing) across both German and English text collections as well as across historical and contemporary data sets (Tables 6-8). What rises to the top here are a host of perceptual categories (seeing, hearing, feeling) that construct the phenomonological reality of an experiencing individual. And the greater prevalence of body words gives us an indication of where that attention most often lies. It is knowledge, not just of otherness, but of another embodied individual that most consistently frames the epistemological horizon of the novel from a quantitative point of view. This result poses an interesting challenge for "theory of mind" approaches that argue that fiction's primary purpose is the enactment of another human consciousness.[29] While we will see an area where this hypothesis does make sense in the next test, in terms of understanding the novel's distinctive fictional qualities the mind-body distinction that underlies theory of mind models does not hold up well in light of the novel's strong emphasis on sensorial input and embodied entities. The sensual experience of a sensing being: this is what appears to be uniquely reiterated in the imaginative work of novelistic writing when compared to its non-fictional counterpart.

Table 6. Third Person Novels v. History

(English 19C, Dialogue Removed)

Feature	Category	3P Novels (%)	History (%)	Ratio	p-value
exclam	linguistic	0,08	0,02	4	***
hear	perception	0,67	0,19	3,53	***
see	perception	1,3	0,37	3,51	***
feel	perception	0,84	0,26	3,21	***
percept	perception	2,95	0,92	3,2	***
shehe	linguistic	6,95	2,25	3,09	***
body	biological	1,27	0,41	3,09	***
sexual	biological	0,19	0,06	3,08	***
assent	oral	0,03	0,01	3	***
qmark	linguistic	0,09	0,03	3	***
home	social	0,73	0,26	2,84	***
anx	affective	0,56	0,24	2,38	***
bio	biological	2,25	0,98	2,3	***

[29] Lisa Zunshine, Why We Read Fiction: Theory of Mind and the Novel (Columbus: Ohio State UP, 2006).

Feature	Category	3P Novels (%)	History (%)	Ratio	p-value
ingest	biological	0,3	0,14	2,14	***
ppron	linguistic	8,42	4,17	2,02	***
sad	affective	0,69	0,36	1,92	***
friend	social	0,17	0,09	1,89	***
family	social	0,45	0,26	1,75	***
discrep	cognitive	1,27	0,74	1,72	***
social	social	12,04	7	1,72	***

Table 7. Third Person Novels v. History
(German 19C, No Dialogue Removed)

Feature	Category	3P Novels (%)	History (%)	Ratio	p-value
I	linguistic	2,16	0,09	24	***
exclam	linguistic	0,57	0,04	14,13	***
self	linguistic	2,61	0,32	8,16	***
qmark	linguistic	0,46	0,06	7,67	***
sexual	biological	0,17	0,03	5,67	***
posfeel	affective	0,43	0,08	5,38	***
see	perception	0,2	0,04	5	***
senses	perception	0,29	0,07	4,07	***
friends	social	0,31	0,08	3,81	***
sleep	biological	0,11	0,03	3,5	***
physical	biological	1,72	0,49	3,5	***
you	linguistic	2,07	0,64	3,23	***
body	biological	1,32	0,41	3,22	***
assent	oral	0,16	0,05	3,2	***
hear	perception	0,03	0,01	3	***
humans	social	0,75	0,27	2,76	***
pronoun	linguistic	11,14	4,05	2,75	***
othref	linguistic	9,31	4,42	2,11	***
we	linguistic	0,4	0,19	2,11	***
family	social	0,72	0,35	2,06	***

Table 8. Third Person Novels v. History
(English Contemporary, Dialogue Removed)

Feature	Category	3P Novels (%)	History (%)	Ratio	P-Value
you	linguistic	0,20	0,03	6,67	***
body	biological	1,84	0,41	4,48	***
qmark	linguistic	0,21	0,05	4,2	***
shehe	linguistic	6,87	2,07	3,32	***
feel	perceptual	1,12	0,34	3,28	***
ingest	biological	0,59	0,2	2,93	***
hear	perceptual	0,7	0,24	2,92	***
percept	perceptual	3,64	1,28	2,84	***
bio	biological	3,1	1,12	2,76	***
see	perceptual	1,44	0,56	2,57	***
home	social	0,94	0,37	2,56	***
assent	oral	0,05	0,02	2,5	***

Feature	Category	3P Novels (%)	History (%)	Ratio	P-Value
ppron	linguistic	9,05	3,84	2,36	***
sexual	biological	0,18	0,08	2,25	***
I	linguistic	0,18	0,09	1,94	***
family	social	0,47	0,26	1,81	***
pronoun	linguistic	13,13	7,6	1,73	***
social	social	12,33	7,25	1,7	***
present	linguistic	2,13	1,47	1,45	***
period	linguistic	6,64	4,66	1,43	***

As a final way to understand, and bring into sharper relief, the significance of what I am calling the phenomenological orientation of the novel, I compare the nineteenth century novel with a particular subset of fiction that excludes novels published during the same time period (HATHI_TALES). Non-novelistic fiction in this case refers to a broad mixture of fictional writing that would have been very present to nineteenth-century readers, including classical epics translated into prose (*The Iliad, Odyssey, Edda, Nibelungenlied*), classic works of prose fiction (*The Tale of Genji, The Decameron*, King Arthur Tales, and Rabelais), fairy tale collections from around the world (drawn from Irish, German, Danish, Japanese and Indian sources), contemporary novella collections (novellas by Hoffmann, Tolstoy, Dickens, Maupassant, Hawthorne, and Washington Irving), as well as a variety of "tales" collections (*Tales of Former Times, Tales of Domestic Life, Moral Tales*). This data set is meant to represent a range of prose fiction that would have been widely read and known to nineteenth-century anglophone readers but would not have been considered a "novel." While the material dates from different epochs, the publications (and translations) are all contemporaneous with the period as a whole.

Table 9. Novels v. Other Fiction

Feature	Category	Novel (%)	Other Fiction (%)	Ratio	P-Value
qmark	linguistic	0,54	0,33	1,65	***
assent	oral	0,17	0,11	1,57	**
period	linguistic	4,92	3,82	1,29	***
present	linguistic	4,45	3,52	1,27	***
discrep	cognitive	1,61	1,31	1,23	***
shehe	linguistic	5,82	4,92	1,18	*
negate	cognitive	1,68	1,43	1,18	***
see	perception	1,16	1	1,16	*
verb	linguistic	12,91	11,17	1,16	***
percept	perception	3,14	2,74	1,15	**
hear	perception	1,12	0,98	1,14	*
insight	cognitive	2,15	1,9	1,13	**
tentat	cognitive	2,27	2,01	1,13	**
auxverb	linguistic	7,78	6,93	1,12	***
past	linguistic	6,2	5,54	1,12	*
feel	perception	0,68	0,61	1,11	,
future	linguistic	1,2	1,09	1,1	*
adverb	linguistic	3,74	3,41	1,1	*
excl	cognitive	2,26	2,09	1,08	*
social	social	13,29	12,37	1,07	*

Three interesting features initially stand out in this table. First, the ratios are much lower when compared with non-fiction. While these groups are similarly well-differentiated when compared to non-fiction, when compared to each other the overall distinctiveness drops considerably. If we run the same classifier as above, we can predict novels with about 68% accuracy, which is close to the threshold of statistical significance (p=0.018). If we use a slightly larger collection of novels from the Hathi Trust collection (428 to mirror "other fiction"), accuracy will increase slightly to 74% (p=7.23e-05). This is still considerably lower for example than the ability to predict novels from different genres. As Ted Underwood has shown,

it is possible to predict detective fiction and science fiction across a 150-year span with between 88-90% accuracy.[30] The broad category of "other fiction" then is not highly differentiated from novels as a subset of fiction.

Second, while we see some of our more familiar linguistic fictional markers such as pronouns and dialogue, we also see a new feature in the category of verbs. There are more verbs overall as well as more varied tenses (past, future, present, in addition to auxiliary verbs). In other words, there appears to be greater temporal complexity to novels than can be found in fiction more generally. While this deserves its own study, it suggests an initial insight into one of the key ways that novels differentiate themselves from other kinds of imaginary writing in the nineteenth century.[31]

Finally, we also see a new category emerge here that we have not seen before, one that falls under the heading of "cognitive process." These are the dictionaries that LIWC labels "discrepancy," "negation," "tentativeness," and "insight." If we examine the words in those dictionaries that are most distinctive of novels (and here I rank by log-likelihood ratio), we can see the extent to which these are words that tend to mark out moments of self-reflection, doubt, and hesitation, a kind of testing-relationship to the world.[32]

Modal verbs in particular are extremely prevalent here, could, would, must, might, and should, as well as their negative contractions, and so too is the act of negation more generally (don't, can't, didn't, not, never, nobody). As the presence of "if" suggests, these groups offer different ways of expressing conditionality or even impossibility. At the same time, indefinite words such as something, somebody, anything and anybody are more prevalent, along with a more specific vocabulary of hesitation (perhaps, chance, hope, possibly, guess, maybe, doubt, uncertain). In between the conditional and the impossible language of the novel, there lies a considerable amount of potentiality - chance, but also skepticism.[33]

Finally, we see how novels are marked by a much stronger use of mental states, captured in major verbs such as know, feel, think, remember, and believe, along with a second layer of less frequent, but similarly distinctive complex cognitive verbs such as admit, ponder, imagine, and forgive (the latter not shown). This is the ground of the novel's reflectiveness, that which binds together doubt and conditionality into a consistent mental state. Indeed, the combination of seem and feel, both of which appear 30% more often in the novel, give us a particular indication of what I am calling the novel's phenomenological orientation. Not the world itself, but a person's encounter with and reflection upon that world - the world's feltness - is what marks out the unique terrain of novelistic discourse when compared with other forms of classical fiction. It is this combination of sense perception plus cognitive skepticism that seems to bring out the novel's contribution to fictional discourse. The novel professes its uniqueness in the way it offers extended reading experiences of the human assessment of the world's givenness.[34]

Table 10. Novels v. Other Fiction - Distinctive Vocabulary

Discrepancy	Frequency (per 100K)	G2	Ratio	Insight	Frequency (per 100K)	G2	Ratio
want	84,11	2086,73	1,77	think	154,93	3241,93	1,68
would	348,14	1534,14	1,25	know	183,74	1564,75	1,38
if	315,85	1213,5	1,23	realiz	5,41	1524,74	31,53
wouldnt	13,15	978,85	3,02	thought	139,88	801,93	1,29
could	246,67	738,47	1,2	seem	120,01	730,36	1,31
couldnt	10,73	683,5	2,72	felt	64,81	557,08	1,38
shouldnt	5,09	638,09	4,86	sens	31,39	547,27	1,6
rather	51,37	472,47	1,39	feel	86,05	483,08	1,29
must	131,82	439,63	1,21	meant	15,03	465,98	1,92
mustnt	2,58	386,01	6,1	recogn	3,78	434,78	4,42

[30] Ted Underwood, "The Life Cycles of Genres," http://culturalanalytics.org/2016/05/the-life-cycles-of-genres/.

[31] This fiction/novel distinction that is translated across the axis of time will recapitulate itself in the genre distinctions of the contemporary novel as they relate to social value. One of the strongest ways bestselling and prizewinning novels in the present differentiate is across the feature of nostalgia and retrospection. See Piper and Portelance, "How Cultural Capital Works."

[32] Results are contained in the table MDW_EN_NOV_3P_v_HATHI_TALES_Final.csv. See Script 4.1.

[33] This observation aligns with Matt Erlin's argument about the philosophical dimensions of novelistic narration. Matthew Erlin, "From the Philosophical to the Epistemic Novel?" CA: Journal of Cultural Analytics (forthcoming).

[34] If we just condition on these four features, the classification results will outperform the average accuracy of four features chosen at random by a statistically significant margin, though to be sure the numbers are considerably lower than when we use all 80 features, and there are other combinations that will perform slightly better. Those feature combinations that do perform better most often contain present tense verbs and categories of sense perception, pointing to the other ways discussed here through which novels are unique. All Features = 76%, Prevarication Freatures = 63.4%, Random4 (100 trials) = 60.8% +- 3.8%.

Discrepancy	Frequency (per 100K)	G2	Ratio	Insight	Frequency (per 100K)	G2	Ratio
hope	68,91	362,06	1,28	conscious	18,16	422,69	1,74
ought	23,95	292,34	1,47	rememb	38,16	413,15	1,44
neednt	2,5	268,92	4,12	believ	57,45	369,91	1,32
should	148,47	175,96	1,12	idea	30,27	331,38	1,44
ideal	3,36	149,13	2,23	knew	65,63	318,91	1,27
wish	65,77	142,12	1,17	question	36,44	267,27	1,34
normal	0,85	136,66	6,74	understand	28,38	235,23	1,37
problem	1,99	112,16	2,53	decid	14,48	197,34	1,51
oughtnt	0,72	105	5,9	mean	64,27	162,91	1,18
need	28,78	99,08	1,22	percept	3,82	122,36	1,94

Negation Frequency (per 100K) G2 Ratio Tentative Frequency (per 100K) G2 Ratio dont 107,1 6643,46 2,68 like 195,81 1262,54 1,32 cant 34,32 2369,73 2,87 someth 62,9 1091,98 1,6 isnt 13,82 1375,74 3,83 quit 74,78 1003,91 1,5 didnt 21,55 1322,7 2,66 ani 162,81 829,51 1,27 doesnt 8,48 961,32 4,35 anyth 43,52 527,88 1,47 havent 7,74 834,73 4,13 vagu 8,38 470,24 2,52 wont 24,89 747,67 1,89 suppos 38 391,83 1,42 not 715,73 559,28 1,1 hope 68,91 362,06 1,28 never 149,5 414,82 1,19 almost 51,55 357,66 1,33 arent 2,41 378,8 6,56 hard 45,05 317,18 1,33 shant 3,19 376,61 4,55 perhap 48,76 311,51 1,31 no 332,1 341,8 1,11 sort 28,85 308,24 1,43 hadnt 5,22 324,28 2,68 might 118,27 269,19 1,17 wasnt 6,94 294,65 2,19 question 36,44 267,27 1,34 noth 86 268,27 1,21 guess 11,43 242,72 1,69 hasnt 2,51 233,4 3,6 possibl 36,26 180,39 1,27 werent 1,25 157,7 4,91 somehow 5,94 179,07 1,9 aint 6,25 58,15 1,4 chanc 19,99 142,57 1,34 nobody 10,45 49,32 1,26 anybodi 7,05 141,16 1,66 negat 1,25 44,62 2,03 anyon 3,27 116,36 2,02

The Question of the Novel's Realism

If we can agree that one of the ways the novel distinguishes itself as a genre is through a more intensive attention to a phenomenological vocabulary, the question arises as to whether such attention is also accompanied by a greater degree of world-focusedness, that is, more attention to reality or what I would more abstractly describe as "givenness" following Lukacs. Are the phenomenological and the realistic mutually exclusive of one another or mutually constitutive? Can we test, in other words, the longstanding hypothesis of the novel's heightened realism?

While these are questions that deserve their own study, I offer two tests here that attempt to gain some insight into the validity of the novel's realistic tendencies. In my first test, I explore the novel's relative attention to abstraction versus physical entities. To do so, I compare a subset of nineteenth-century novels from the Chadwyck Healey collection (about 700 novels) with my "other fiction" from the Hathi Trust and translate these collections into their respective hypernym trees using Wordnet.[35] Hypernyms provide higher-order classifications of nouns ("furniture" is a hypernym of "chair," for example) and can allow us to see whether particular categories are more present than others in a given corpus, much in the way LIWC does for emotional and psychological processes. So for example, if the word "marsh" appears in a novel, this would be translated into "wetland," "land," "ground," "soil," "object," "physical_object," "physical_entity," and "entity." All nouns in this case are "entities," while their first order of distinction is between those nouns that are "physical," like a marsh, and those that are abstract, like "death."

The question that this model allows us to pose is whether novels exhibit significantly higher amounts of physical entities when compared with other kinds of classical fiction (or histories for that matter). By doing so, we can gain confidence as to whether the novel's uniqueness is tied to its physical objectivity, one potential way of understanding its degree of realism, or whether it hinges more on abstract mental or emotional states.

The second test attempts to understand realism as a greater amount of specificity. The more specific a text is, the more focused it is on the world around it. To test this, I compare the percentage of words in one text that are hypernyms of another text's words (and vice versa). The more words from one group that are hypernyms of words from another, the

[35]The novels are first reduced to only their nouns using openNLP in R and then the transcription is run using the python script hypernyms_ReturnTopClass.py.

more abstract that first group can be said to be and the more specific the second group. For example, if I use the word "marsh" and you use the word "land," then my language can be said to be more specific than yours (and yours more abstract than mine). Unlike in my first test, where having more objects in a text is a way of thinking about an attention to realism, here the emphasis is on measuring a greater degree of specificity as a marker of the real.

The results suggest that the novel's relationship to its concreteness, when measured in this way, has indeed been changing over the course of the nineteenth century (Table 11).[36] If we look at the first half of the nineteenth century, we see how there is a greater degree of abstraction relative to physical objects when compared to classical fiction and tales, but that this difference disappears by the second half of the century. As Ryan Heuser and Long Le Khac have argued, the British novel experiences a decline of valuation and a rise of concretization over the course of the century.[37] And yet an important caveat to that finding is that while abstraction appears to be declining in the novel, it never drops below other kinds of fictional discourse from the period. Far from the Victorian novel being uniquely concrete, it is the early-nineteenth-century novel that looks uniquely abstract when compared to other kinds of fictional discourse from the period. It suggests that we have been potentially telling this story in reverse: what matters in the nineteenth century is not the later rise of concreteness, which looks more like other types of fictional writing, but the earlier abstractness of the novel, which stands out relative to other types of fiction (not to mention abstraction remains considerably more important to these texts overall than their physicality).

The other part of the table, that which concerns the novel's specificity, tells this same story in the other direction. Where the first half of the century witnessed little significant difference in the degree of specificity between the novel and other kinds of fiction (about a 0.003% difference), by the second half the century, the novel has about 0.5% more specific words than other kinds of fiction (or about 2 words/page).[38] In other words, the novel approaches other kinds of fiction in its degree of abstraction while it departs from other kinds of fiction in its degree of specificity. It begins the century uniquely abstract and finishes uniquely specific.

There are of course numerous other ways we might think about the novel's realism. But these initial results suggest that the Wattian thesis about the novel's realism, understood here as a greater degree of both physicality and specificity, appears, in the first instance, less as a story of exceptionality and more as a regression to the norms of fictional discourse more generally. In the second instance, where the novel does become exceptional in its specificity, this occurs much later than has been traditionally argued. The novel's earlier quantitative rise in the late eighteenth and early nineteenth century appears to be marked instead by a higher deegree of conceptual sophistication and generality, the novel's love of abstraction, far more than its presentation of the world around it.[39] As Matthew Erlin has argued, there is a philosophical dimension to the novel that is an important part of its history that we have so far overlooked.[40] The scholarly emphasis on the realist novel's concretization has missed one of the primary ways through which novels have historically mattered as a form of writing, i.e. in their abstractness.

[36]The data for this table can be found here: Abstraction_Ratio_English750_1800_1849_NounsOnly.csv, Abstraction_Ratio_English750_1850_1899_NounsOnly.csv, Abstraction_Ratio_Hathi_Tales_NounsOnly.csv, Hypernymy_English_Tales_v_750_1_TakeAll.csv, Hypernymy_English_Tales_v_750_2_TakeAll.csv, Hypernymy_English750_1_v_Tales_TakeAll.csv, Hypernymy_English750_2_v_Tales_TakeAll.csv. See Script 5.1 and 5.2 for a fuller explanation of how the scores were calculated.

[37]Ryan Heuser and Long Le Khac, "A Quantitative Literary History of 2,958 Nineteenth-Century British Novels: The Semantic Cohort Method," Stanford Literary Lab Pamphlet 4: http://litlab.stanford.edu/LiteraryLabPamphlet4.pdf.

[38]One of the reasons for the extremely low p-value here is the high number of observations. Given that we are comparing every novel to every other novel across groups, we get over 13,000 observations. Even small differences can look significant. The important point is the actual difference in this case which changes from 0.003% to 0.5%.

[39]This understanding of the novel's asbtractness as key to its earlier distinctiveness offers a different way of thinking about a critique of the realist hypothesis than, say, Wayne C. Booth. For Booth, the novel's significance, or for him the "good" novel's significance, is its ability to focus on dramatic intensity rather than realism. "The interest in realism is not a 'theory' or even a combination of theories that can be proved right or wrong; it is an expression of what men [sic] at a given time have cared for most." This brings us closer to a theory of plot as arc in the spirit of Matthew Jockers' work, the rhythmic rise and fall of emotional intensities. See Wayne C. Booth, The Rhetoric of Fiction (Chicago UP, 1983), 63.

[40]Matthew Erlin, "From the Philosophical to the Epistemic Novel?" CA: Journal of Cultural Analytics (forthcoming).

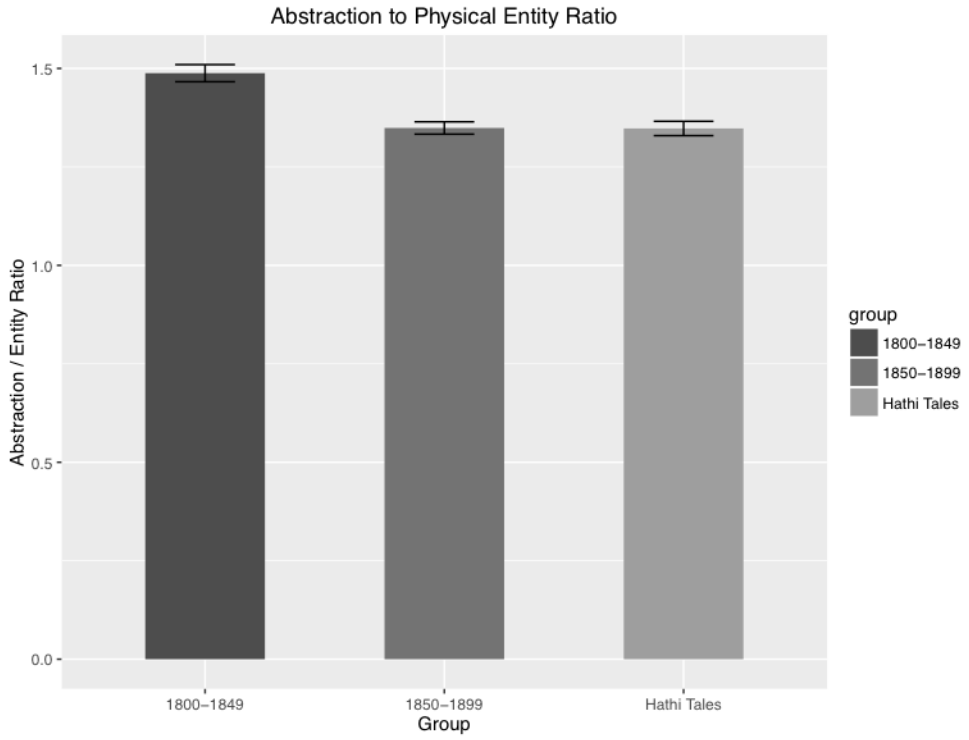


Figure 1. This graph measures the ratio of nouns identified as "abstract" versus "physical entity" by WordNet classification.

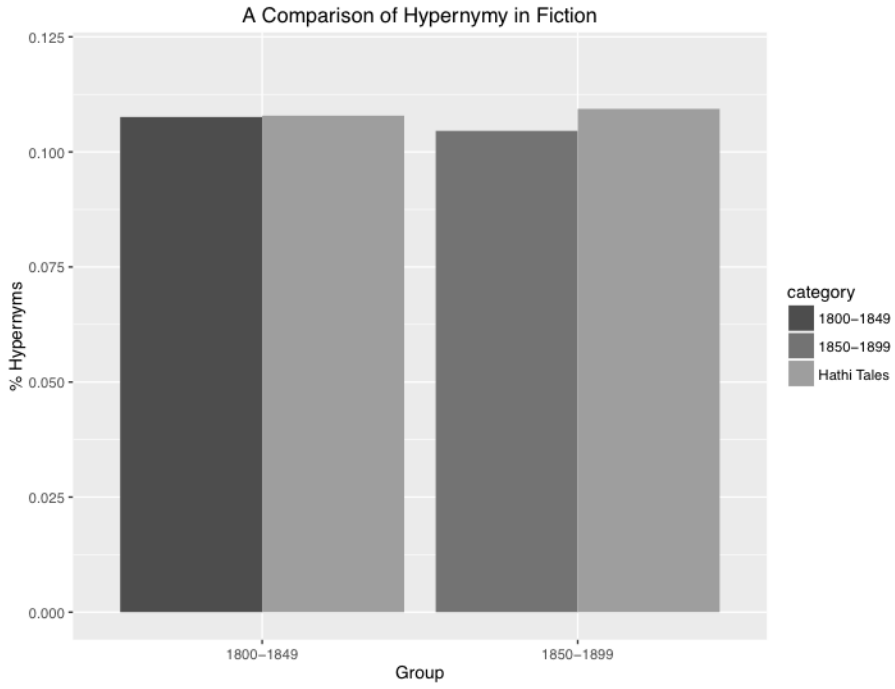
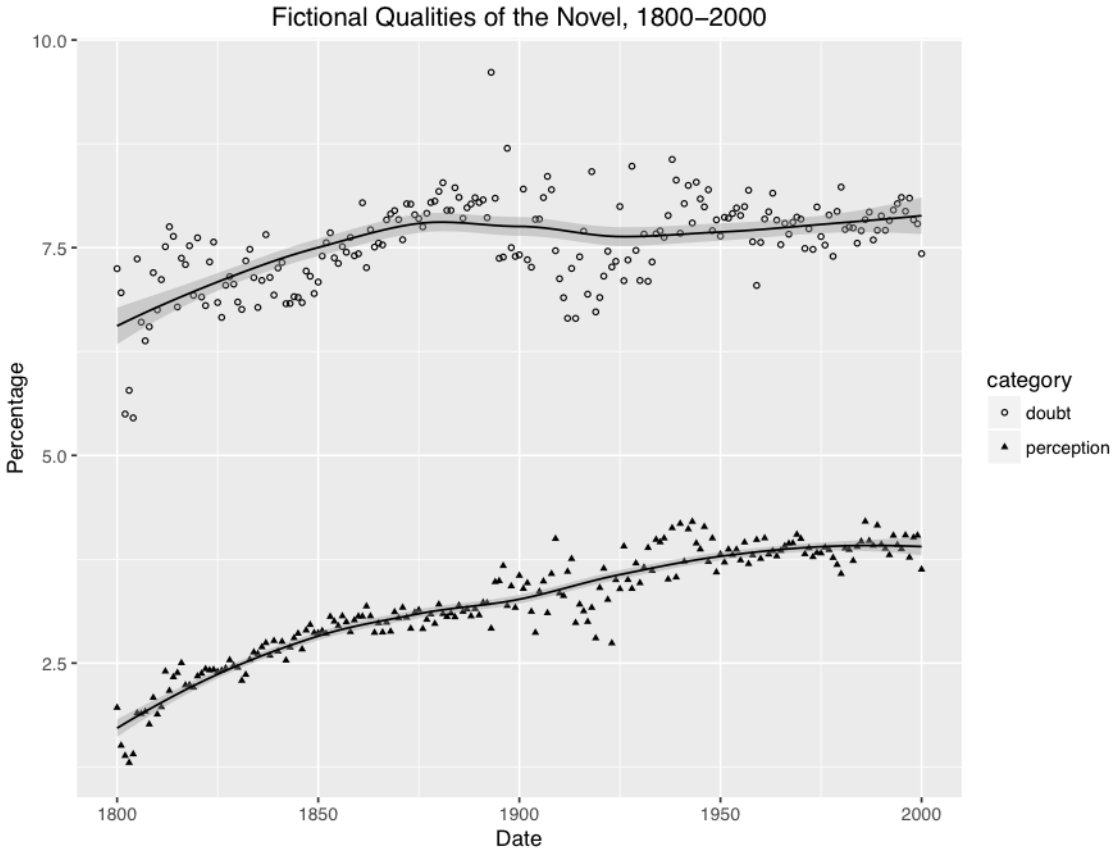


Figure 2. This graph measures the percentage of words in one corpus that are hypernyms of words in the other corpus. A hypernym is a more abstract representation than its hyponym. Error bars have been removed because they are too narrow to visualize.

Conclusion

In trying to distinguish fiction from non-fiction, in locating what makes fiction and the novel unique as types of writing, I have in the process been attempting to gain insights into their larger social function, to answer that perennial question of "why literature matters." According to the results presented here, if we focus on the quantitatively distinct qualities of novels in particular - of what separates them off from non-fictional or "true" writing - we can say that the novel's mattering since the nineteenth century appears to be less a matter of social realism and more one of phenomenological encounter, a kind of social imbedding in the world. It is not that this is the only way novels have been or could be meaningful to readers. This is the problem of predictive versus descriptive approaches that I discussed at the outset - predictability forecloses other possibilities, other qualities that may be important. Descriptive models simply identify which features differ and by how much, without presupposing a limit to the feature space. The value of the quantitive point of view is that it allows us to better understand the way a particular type of writing signals to readers a particular orientation, what we might call its social positionality (via Bourdieu). This does not foreclose the myriad ways readers can find their own version of the novel's mattering. But it does allow us to better understand the novel as a social category.

Seen in this way, the fictionality of novels is special because of the notion of encounter and the questioning that comes with it, the way they put us as readers in the world in a particular way. Things seem and feel a certain way, just as there is a great deal of doubt and chance and perhaps and maybe. These findings appear to be robust across two different languages, two very different time frames, and across both a larger and smaller, more canonical sample of writing. As we have already seen, we can use models built on the behavior of the classical nineteenth-century novel and still predict contemporary novels with a great deal of accuracy. If we look more closely at these features over time, we also see how they both increase and then eventually remain constant over a longer period of time (fig. 1). The values that are put in place surrounding fictional discourse in the nineteenth century remain largely intact over the course of time, with some features, like an emphasis on sense perception rising considerably in importance.[41]



[41] Further analysis suggests that this increase is largely driven by a reliance on words for "site." Most of the other senses remain flat. More research could uncover what this ocular bias suggests in terms of fiction.

Figure 3. Rates of "doubt" and sense "perception" in English novels published between 1800 and 2000. In the former case we see a very slight rise by the second half of the nineteenth century and in the latter a continual rise that levels off in the postwar era.

Of course not all novels are like this and not all novels that are like this are like this in equal ways. The extent to which these markers of fictionality are consistent across different genres of novelistic writing still remains an open question. One can also imagine another study in which we could explore those moments when novels become highly non-fictional, to understand the truth within the imaginary. How might we characterize the non-fictional within the fictional and what purpose do these passages perform? It will come as little surprise that of the handful of novels misclassified by the algorithm used at the beginning of this article, Melville's Typee and Omoo are in this group (but not Moby Dick, suggesting canonization is picking-up at least in part on its fictionality). But so too is a novel like Behemoth: A Legend of the Mound-Builders by Cornelius Matthews, a novel about slaying an ancient Mastodon and argued to be a key influence on Melville.[42] Can we say with more specificity what function the informational use of language has within fiction - Aristotle's legein - not only in these novels, but within novels more generally? Such a study would try to provide a mirror image to this article's insights, a diptych in which we see something about the novel reflected in relief.

Most importantly for me though is the way the methods used here are not reducible to the single passage or well-turned sentence. There are hundreds of thousands of lines of novels that contain instances of our fictional markers, of pronouns, sense perception, and the subjunctive mood and negation. Each one of them is slightly different from another. Individually they are nothing special. Taken together, however, they signal a powerful message to readers. Fictionality is a feeling we get as readers from the likelihood of seeing all of these words flash across the page, words like "feeling," "knowing," "seeing," "remembering," "almost," "possibly," "vaguely," and a variety of forms of negation, of the not. This is the space of fiction's apartness. "And where is this elsewhere?" Roland Barthes once asked. "In the paradise of words."

[42] Curtis Dahl, "Moby Dick's Cousin Behemoth," American Literature 31.1 (1959), 21-29.